Detection Signal Design for Failure Detection: a Robust Approach

Ramine Nikoukhah, Stephen L. Campbell, François Delebecque

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Detection signal design for failure detection: a robust approach

Ramine Nikoukhah , Stephen L. Campbell , François Delebecque

No 3547

Novembre 1998

THÈME 4

ISSN 0249-6399

ROCQUENCOURT

Detection signal design for failure detection: a robust approach

Ramine Nikoukhah , Stephen L. Campbell* , François Delebecque

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet META2

Rapport de recherche n ° 3547 — Novembre 1998 — 34 pages

Abstract: The detection signal is an input signal that enhances the detectability of failure. Assuming that the normal and the failed behaviors of a process can be modeled by two linear systems subject to bounded energy perturbations, a method for constructing a minimum energy detection signal, guaranteeing detection of failure, is presented. The online implementation of the failure decision filter is also considered.

Key-words: Failure detection, detection signal, descriptor systems, robustness.

(Résumé : tsvp)

* Dept. of Mathematics, North Carolina State University, Raleigh, NC 27695-8205. USA. Research supported in part by the National Science Foundation under ECS-9500589, DMS-9423705, INT-9605114, and DMS-9802259.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : (33) 01 39 63 55 11 – Télécopie : (33) 01 39 63 53 30

Conception de signal de détection pour la détection de pannes : une approche robuste

Résumé : Un signal de détection est un signal d'entrée qui permet de détecter au mieux une occurence de pannes. On présente une méthode de construction d'un signal de détection d'énergie minimale garantissant la détection. On suppose que les comportements, normal ou défaillant, peuvent être modélisés par deux systèmes linéaires soumis à des perturbations d'énergie bornée. On considère aussi l'implémentation en temps réel du filtre détecteur.

Mots-clé : Détection de pannes, signal de détection, systèmes implicites, robustesse.

1 Introduction

There are two approaches to the problem of failure detection and isolation. The first is a passive approach where the detector monitors the inputs and the outputs of the system and decides whether (and if possible what kind of) a failure has occurred. This is done by comparing the measured input-output behavior with "normal" behavior of the system. The passive approach is used to continuously monitor the system in particular when the detector has no way of acting upon the system, for material or security reasons. Most of the work in the area of failure detection is geared towards this type of approach [1, 10, 13].

The active approach to failure detection consists in acting upon the system on a periodic basis or at critical times using a test signal, which we call a detection signal, in order to exhibit abnormal behaviors which would otherwise remain undetected during normal operation. The detector can act by taking over part or all of the inputs of the system for a period of time: the test period. The decision whether or not the system has failed should be made at (and if possible before) the end of the test period. The structure of the failure detection method considered in this paper is depicted in Figure 1.1 .

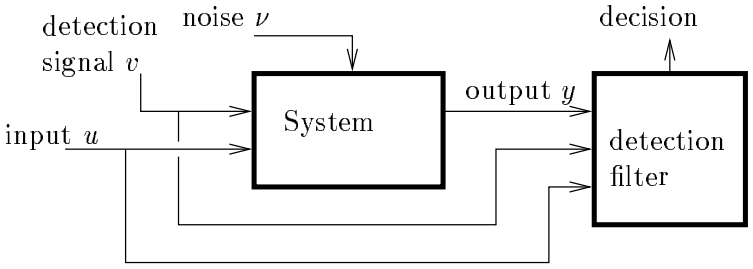

Figure 1.1 : Active failure detection.

The design of detection signals has been a major issue in system identification but their use for failure detection has been introduced in [14, 5, 6]; see also [11]. The detection signal (called auxiliary input) in these works are considered to be linear inputs of stochastic models and their objective is to optimize some statistical properties of the detector. In [8], a method for guaranteed failure detection was presented in which perturbations were modeled as polyhedral sets. The method was based on solving large linear programming problems. The work presented here is closely related to that work but the approach is different in that uncertainties are modeled as bounded energy signals.

Let $v = \{v(k)\}$, $k \in [0, N-1]$, denote a detection signal and let $\mathcal{A}^0(v)$ represent the set $\{u, y\}$ of normal input-output behaviors of the system over a period of time of length N; $y = \{y(k)\}$ and $u = \{u(k)\}$, $k \in [0, N-1]$, denote respectively the measured output vector sequence and the measured input vector sequence. Similarly, let $\mathcal{A}^1(v)$ represent system behavior when failure has occurred. Then failure detection consists of observing the inputs and outputs of the system over some period of time of length N, called the detection horizon, and deciding to which set they belong. Clearly for perfect isolation we need that for a given detection signal v

$$\mathcal{A}^0(v) \cap \mathcal{A}^1(v) = \emptyset. \tag{1.1}$$

In this paper, we focus on the following problem: how to construct a detection signal v so that condition (1.1) holds. The solution to this problem can be computationally intensive since

it is constructed, once and for all, in advance. We also briefly consider the problem of (assuming condition (1.1) holds) how to decide to which set the observations belong to. The solution to this problem is to be implemented on-line and needs to be efficient.

The outline of the paper is as follows. In Section 2, basic assumptions are presented and the model is introduced. Detection signals are characterized in Section 3 and a method for constructing an optimal solution over finite horizon is given. The asymptotic properties of the detection signal as the horizon goes to infinity are studied in Section 4 where it is shown that for long detection horizons, asymptotically optimal detection signals can be easily characterized and constructed. Section 5 is dedicated to a brief discussion of the implementation of on-line decision filters.

2 System model

The behaviors of the normal and failed modes of the system are supposed to be modeled by

$$\begin{aligned} x_i(k+1) &= A_i x_i(k) + B_i u(k) + D_i v(k) + M_i \nu_i(k), & (2.1)\\ y(k) &= C_i x_i(k) + N_i \nu_i(k) & (2.2) \end{aligned}$$

for $k = 0, \ldots, N-1$, and $i = 0$ and 1 correspond respectively to the normal and failed modes. v is the detection signal (known), u and y are inputs and outputs which are measured on-line. A_i, B_i, C_i, D_i, M_i, N_i are matrices of appropriate dimensions. Note that x_0 and x_1 need not even have the same dimension. The same is true for ν_0 and ν_1.

The two models are supposed to satisfy:

$$\begin{pmatrix} zI - A_i & M_i \\ C_i & N_i \end{pmatrix} \quad \text{has full row rank } \forall z, \tag{2.3}$$

and N_i has full row rank. These assumptions are the usual assumptions for Kalman filtering problems. For example when the dynamic and observation noises are independent, i.e., $M_i N_i^T = 0$, Assumption (2.3) is equivalent to the controllability of (A_i, M_i).

Finally, the $\nu_i(k)$'s are unknown sequences representing perturbations, noises and unmeasured inputs. They are supposed to satisfy

$$\|\nu_i\|^2 = \sum_{k=0}^{N-1} \|\nu_i(k)\|^2 < \sigma, \quad i = 0, 1. \tag{2.4}$$

More general constraints such as $\sum \nu_i(j)^T R_i \nu_i(j) < \sigma_i$ can always be converted to (2.4) by modification of the M_i and N_i matrices.

Note that unlike most other approaches to uncertainty modeling in dynamical systems for the purposes of failure detection, ν is not a stochastic white noise sequence, but rather a bounded energy arbitrary discrete sequence. This kind of modeling is used in robust control methodology (in particular H_∞), and is particularly useful when there is large uncertainty concerning the power density of the noise process.

Finally, a fundamental assumption, in this paper, is that, during the test period, the system is either in normal mode or failed mode. No transition occurs during the test period.

3 Detection signal design

3.1 Proper detection signal

We say that v is a *proper detection signal* if its application implies that we are able to always distinguish the normal mode from the failed mode of the system, based on observations u and y.

Definition 3.1 *Detection signal v is not proper if there exist sequences x_0, x_1, ν_0, ν_1, u and y satisfying (2.1), (2.2) and (2.4) both for $i = 0$ and $i = 1$. The detection sequence v is called proper otherwise.*

The objective here is to find a method for constructing a minimum energy proper detection signal. For that, we need first to introduce an auxiliary cost function.

Definition 3.2 *The function $J(\beta, v)$ is the auxiliary cost function associated with problem (2.1)-(2.2) if*

$$J(\beta, v) = \min_{\substack{x_i, \nu_i, u, y \\ \text{subject to (2.1)-(2.2), } i=0,1}} \sum_{k=0}^{N-1} \beta \|\nu_0(k)\|^2 + (1-\beta)\|\nu_1(k)\|^2 \tag{3.1}$$

for $0 \leq \beta \leq 1$.

Note that ν_0, ν_1 in (3.1) need not satisfy (2.4).

Lemma 3.1 *For all v, for $0 \leq \beta \leq 1$, $J(\beta, v)$ is defined and has the following properties:*

1. *it is zero for $\beta = 0$ and $\beta = 1$,*
2. *it is quadratic in v, i.e., for all scalar c, $J(\beta, cv) = |c|^2 J(\beta, v)$.*
3. *it is a continuous function of β,*
4. *it satisfies $J(\beta, v) < \sigma$ if v is not proper,*
5. *it is a strictly concave function of β if the set of proper detection signals is not empty, otherwise it is identically zero.*

Proof $J(\beta, v)$ is well defined for all v thanks to full rankedness of N_i's since this clearly implies that the minimization in (3.1) is over a non-empty set for all v.

Let $\nu_0 = 0$ and consider constraints (2.1)-(2.2) for $i = 0$. Clearly there exist x_0, u and y so that these constraints are satisfied. But then for these u and y (and in fact for any other), there exist x_1 and ν_1 satisfying constraints (2.1)-(2.2) for $i = 1$. This shows that $\nu_0 = 0$ is consistent with (2.1)-(2.2) for $i = 0$ and $i = 1$. Thus $J(1, v) = 0$. Similarly, we can show that $\nu_1 = 0$ is consistent and $J(0, v) = 0$. Thus, $J(0, v) = J(1, v) = 0$ for all v. This proves the first statement.

Note that (2.1), (2.2) for $i = 0, 1$ and $k = 0, \ldots, N-1$ form a consistent set of linear equations in the variables $\{x_0, x_1, \nu_0, \nu_1, u, y\}$ which have a right hand side of the form Qv. Thus the set of all solutions is an affine set of the form $\widehat{Q}v + \sum_j \alpha_j z_j$ for some vectors z_j in $\{x_0, x_1, \nu_0, \nu_1, u, y\}$ space.

Projecting this affine set onto the $\{\nu_0, \nu_1\}$ components we get the the set of pairs (ν_0, ν_1) for which (2.1) and (2.2) hold has the form $\overline{Q}v + \sum_j \alpha_j \overline{z}_j$ for some vectors $\overline{z}_j$ in $\{\nu_0, \nu_1\}$ space. Equivalently, there are matrices $\mathcal{A}, \mathcal{B}, \mathcal{Q}$ so that the consistent ν_i are characterized by the system

$$\mathcal{A}\nu_0 + \mathcal{B}\nu_1 = \mathcal{Q}v \tag{3.2}$$

This system is consistent for all v so that we may assume without loss of generality that $\begin{pmatrix}\mathcal{A} & \mathcal{B}\end{pmatrix}$ has full row rank. Let

$$V_\beta = \begin{pmatrix} \beta I & 0 \\ 0 & (1-\beta)I \end{pmatrix} \tag{3.3}$$

where the identities are the sizes of ν_0 and ν_1 respectively. Then $J(\beta, v)$ is the minimum, over consistent (ν_0, ν_1), of $\begin{pmatrix}\nu_0^T & \nu_1^T\end{pmatrix} V_\beta \begin{pmatrix}\nu_0 \\ \nu_1\end{pmatrix}$ which we denote $\|\nu\|_\beta^2$ and call the β-norm. We see that $J(\beta, v)$ is the square of the β-norm of the β-norm least squares solution of (3.2). Formulas for weighted least squares solutions can be found in [2]. Translating these back into the Euclidean norm, we get that, for $0 < \beta < 1$,

$$J(\beta, v) = \| \left[\begin{pmatrix}\mathcal{A} & \mathcal{B}\end{pmatrix} V_\beta^{-1/2} \right]^\dagger \mathcal{Q}v \|^2. \tag{3.4}$$

This proves Statement 2.

The matrix $\begin{pmatrix}\mathcal{A} & \mathcal{B}\end{pmatrix} V_\beta^{-1/2}$ is continuous, differentiable and constant rank for $0 < \beta < 1$. Thus its Moore-Penrose inverse is continuous and differentiable. This proves Statement 3 (continuity of $J(\beta, v)$ is obvious at 0 and 1). Note that we get more than continuity, in particular, we get that for fixed v, $J(\beta, v)$ is real analytic (has power series expansions) for $0 < \beta < 1$.

For the fourth statement, note that if v is not proper, then there exist consistent ν_i with $\|\nu_i\|^2 < \sigma$. Thus $J(\beta, v) < \sigma$ for all β. In fact, $J(\beta, v)$ is bounded by the larger of the two $\|\nu_i\|^2$.

Finally, let $0 \le \alpha \le 1$, and $0 \le \beta_1 < \beta_2 \le 1$, then

$$\alpha J(\beta_1, v) + (1-\alpha) J(\beta_2, v) = \alpha \min \sum_{k=0}^{N-1} \beta_1 \|\nu_0(k)\|^2 + (1-\beta_1)\|\nu_1(k)\|^2 + (1-\alpha) \min \sum_{k=0}^{N-1} \beta_2 \|\nu_0(k)\|^2 + (1-\beta_2)\|\nu_1(k)\|^2 \tag{3.5}$$

which is less than or equal to

$$\min \sum_{k=0}^{N-1} (\alpha\beta_1 + (1-\alpha)\beta_2)\|\nu_0(k)\|^2 + (\alpha(1-\beta_1) + (1-\alpha)(1-\beta_2))\|\nu_1(k)\|^2 = J(\alpha\beta_1 + (1-\alpha)\beta_2, v) \tag{3.6}$$

which implies that $J(\beta, v)$ is concave in β. But we have shown that $J(\beta, v)$ is also real analytic, which implies, thanks to Statement 1, that J is either strictly concave or identically zero. Statement 5 follows then thanks to Statement 4. ■

Theorem 3.1 *Suppose the set of proper detection signals is not empty. Let* (β^*, w^*) *be a solution of*

$$\gamma^* = \max_{\substack{0 \le \beta \le 1 \\ w \neq 0}} \frac{J(\beta, w)}{\|w\|^2}. \tag{3.7}$$

Then, a minimum energy detection signal is given by

$$v = \sqrt{\frac{\sigma}{\gamma^*}} \frac{w^*}{\|w^*\|}. \tag{3.8}$$

Proof From Statement 2 of Lemma 3.1, we have that

$$\frac{J(\beta, w)}{\|w\|^2} = J(\beta, \frac{w}{\|w\|}) \tag{3.9}$$

Thus the max in (3.7) is the max of a continuous function $J(\beta, z)$ over a compact set consisting of $0 \le \beta \le 1$, $\|z\| = 1$ and the max (3.7) is attained at least at one place. Note that if v is given by (3.8), then $J(\beta^*, v) = \|v\|^2 \gamma^* = \sigma$. Thus v is proper by Statement 4 of Lemma 3.1.

We now show that v is minimal. Suppose there exists a proper detection signal $\hat{v}$ such that

$$\|\hat{v}\|^2 < \|v\|^2 = \sigma/\gamma^* \tag{3.10}$$

and let

$$\hat{\gamma} = \max_{\beta} \frac{J(\beta, \hat{v})}{\|\hat{v}\|^2} \tag{3.11}$$

and denote by $\hat{\beta}$ a maximizing β in (3.11). Note that $\hat{\gamma} \le \gamma^*$. Note also that

$$J(\hat{\beta}, \hat{v}) = \min \sum_{k=0}^{N-1} \hat{\beta} \|\nu_0(k)\|^2 + (1 - \hat{\beta}) \|\nu_1(k)\|^2 \tag{3.12}$$

where the minimization is subject to (2.1)-(2.2), $i = 0, 1$. Suppose $(\hat{x}_i, \hat{\nu}_i, \hat{u}, \hat{y})$ is a solution of the minimization in (3.12). Then, as we shall show below,

$$\sum_{k=0}^{N-1} \|\hat{\nu}_0(k)\|^2 = \sum_{k=0}^{N-1} \|\hat{\nu}_1(k)\|^2 \tag{3.13}$$

which implies that

$$J(\hat{\beta}, \hat{v}) = \hat{\gamma} \|\hat{v}\|^2 \le \gamma^* \|\hat{v}\|^2 < \sigma. \tag{3.14}$$

But then $(\hat{x}_i, \hat{\nu}_i, \hat{u}, \hat{y})$ satisfy (2.1), (2.2) and (2.4), simultaneously for $i = 0$ and $i = 1$. Thus $\hat{v}$ is not proper which is a contradiction. So, v is a minimum energy proper detection signal.

What remains to be shown is (3.13). This is done first by noting that ν_0 and ν_1 minimizing in (3.1) are unique and continuous functions of β, for all v (see the proof of Lemma 3.1). Since for $\beta = 0$ the corresponding ν_1 is zero, and for $\beta = 1$, ν_0 is zero, there exists a $0 \le \bar{\beta} \le 1$ for which the corresponding optimal ν_i's, denoted $\bar{\nu}_i$, satisfy

$$\|\bar{\nu}_0\|^2 = \|\bar{\nu}_1\|^2. \tag{3.15}$$

So all we need to show is that for $v = \hat{v}$, $\bar{\beta} = \hat{\beta}$. Note that by definition (since $\hat{\nu}_i$'s are minimizing values corresponding to $\hat{\beta}$),

$$\hat{\beta}\|\hat{\nu}_0\|^2 + (1-\hat{\beta})\|\hat{\nu}_1\|^2 \leq \hat{\beta}\|\bar{\nu}_0\|^2 + (1-\hat{\beta})\|\bar{\nu}_1\|^2 \tag{3.16}$$

which thanks to (3.15) implies that

$$\hat{\beta}\|\hat{\nu}_0\|^2 + (1-\hat{\beta})\|\hat{\nu}_1\|^2 \leq \bar{\beta}\|\bar{\nu}_0\|^2 + (1-\bar{\beta})\|\bar{\nu}_1\|^2. \tag{3.17}$$

But by definition, $\hat{\beta}$ is a maximizing β, thus

$$\bar{\beta}\|\bar{\nu}_0\|^2 + (1-\bar{\beta})\|\bar{\nu}_1\|^2 \leq \hat{\beta}\|\hat{\nu}_0\|^2 + (1-\hat{\beta})\|\hat{\nu}_1\|^2 \tag{3.18}$$

which implies that both sides of (3.18) must be equal, i.e., $J(\bar{\beta}, \hat{v}) = J(\hat{\beta}, \hat{v})$. But thanks to Statement 5 of Lemma 3.1, J is strictly concave, thus $\bar{\beta} = \hat{\beta}$. ■

Note that γ^* defined in (3.7) can be considered as a measure of how easy it is to distinguish the two models. The larger γ^* is, the easier it is to separate the two models. And when $\gamma^* = 0$, then the two model are indistinguishable no matter what the input v is. So, γ^* can be considered as the counter part of the Kullback distance [7] used in some stochastic formulations.

Definition 3.3 *We call $\sqrt{\gamma^*}$ the separability index where γ^* is defined in (3.7).*

The optimization problem (3.7) can be expressed as follows:

$$\gamma^* = \max_{0 \leq \beta \leq 1} J^*(\beta) \tag{3.19}$$

where

$$J^*(\beta) = \max_{v \neq 0} \frac{J(\beta, v)}{\|v\|^2}. \tag{3.20}$$

The usefulness of (3.8) depends on having a good estimate of γ^*. The optimization problem (3.19) is a scalar problem over a finite interval. Even though $J^*(\beta)$ is not concave, it has nice properties which make the optimization problem (3.19) not so difficult to solve numerically. In particular, thanks to Lemma 3.1, we know that $J^*(\beta)$ is a max over concave functions each of which is zero at $\beta = 0$ and $\beta = 1$. Using this fact, it is easy to show the following result.

Lemma 3.2 *Consider two scalars β_1 and β_2 satisfying $0 \leq \beta_1 < \beta_2 \leq 1$. Then*

$$\max_{\beta_1 \leq \beta \leq \beta_2} J^*(\beta) \leq \frac{J^*(\beta_1)J^*(\beta_2)}{J^*(\beta_1)(1-\beta_2) + J^*(\beta_2)\beta_1}. \tag{3.21}$$

The proof follows a straightforward geometric argument and is illustrated in Figure 3.1.

Now consider the following simple optimization strategy for estimating γ^* which consists of taking the maximum of $J^*(\beta)$ for $n-1$ regularly spaced values of β over $[0, 1]$:

$$\hat{\gamma} = \max_{k=1,..,n-1} J^*(k/n). \tag{3.22}$$

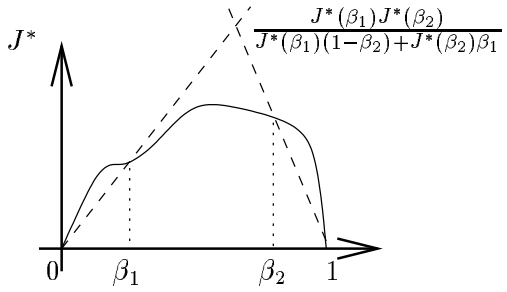

Figure 3.1 : Since $J^*(\beta)$ is the max over concave functions going through the points $(0,0)$ and $(1,0)$, it remains necessarily below the two dashed lines inside $[\beta_1, \beta_2]$ (and above outside).

Then, thanks to Lemma 3.2, it is straightforward to show that

$$\frac{\gamma^* - \hat{\gamma}}{\hat{\gamma}} \leq \frac{1}{n-1}. \tag{3.23}$$

This shows that we are not dealing with a difficult optimization problem. We can of course use more sophisticated algorithms to estimate γ^*.

The next problem is the construction of the auxiliary cost function $J(\beta, v)$ and $J^*(\beta)$. In theory at least, for constructing $J(\beta, v)$, we can convert the equations of the two dynamical systems into a single equation by stacking up both system equations over $[0, N]$, and constructing an explicit solution to the linear quadratic optimization problem (3.1), as it was done in the proof of Lemma 3.1. This however requires manipulating (and in particular inverting) huge matrices when N and the sizes of the states of our two models are large. In the next section, we show that the solution to this problem can be constructed recursively. This can be numerically more efficient, but the real reason for studying this problem is that it allows us to study the asymptotic behavior of the solution as N goes to infinity. And, in particular, find a simple method for constructing detection signals over long horizons.

3.2 Construction of $J(\beta, v)$ and $J^*(\beta)$

The construction of the auxiliary cost function $J(\beta, v)$ can be done recursively. As we have seen in the previous section, the optimization problem to solve is the following:

$$J(\beta, v) = \min \sum_{k=0}^{N-1} \beta \|\nu_0(k)\|^2 + (1-\beta)\|\nu_1(k)\|^2 \tag{3.24}$$

subject to

$$\begin{aligned}
x_0(k+1) &= A_0 x_0(k) + B_0 u(k) + D_0 v(k) + M_0 \nu_0(k) && (3.25)\\
y(k) &= C_0 x_0(k) + N_0 \nu_0(k) && (3.26)\\
x_1(k+1) &= A_1 x_1(k) + B_1 u(k) + D_1 v(k) + M_1 \nu_1(k) && (3.27)\\
y(k) &= C_1 x_1(k) + N_1 \nu_1(k) && (3.28)\\
& && (3.29)
\end{aligned}$$

This problem can be expressed as follows:

$$J(\beta, v) = \min \sum_{k=0}^{N-1} \nu(k)^T V_\beta \nu(k) \tag{3.30}$$

subject to

$$E\xi(k+1) = F\xi(k) + G\nu(k) + Hv(k), \tag{3.31}$$

where

$$E = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & I & 0 \end{pmatrix}, \quad F = \begin{pmatrix} A_0 & 0 & 0 & B_0 \\ 0 & A_1 & 0 & B_1 \\ C_0 & 0 & 0 & 0 \\ 0 & C_1 & 0 & 0 \end{pmatrix}, \tag{3.32}$$

$$G = \begin{pmatrix} M_0 & 0 \\ 0 & M_1 \\ N_0 & 0 \\ 0 & N_1 \end{pmatrix}, \quad H = \begin{pmatrix} D_0 \\ D_1 \\ 0 \\ 0 \end{pmatrix}, \quad V_\beta = \begin{pmatrix} \beta I & 0 \\ 0 & (1-\beta)I \end{pmatrix}, \tag{3.33}$$

and

$$\nu(k) = \begin{pmatrix} \nu_0(k) \\ \nu_1(k) \end{pmatrix}, \quad \xi(k) = \begin{pmatrix} x_0(k) \\ x_1(k) \\ y(k) \\ u(k) \end{pmatrix}. \tag{3.34}$$

Constraints (3.31) can be simplified without affecting the solution of the optimization problem (3.30). For example, since $u(k)$ and $y(k)$ are not dynamical variables and they do not appear in the cost function, they can be removed from the constraints by simple matrix operations. In particular, it suffices to replace (3.26) and (3.28) with their difference (in which $y(k)$ does not appear), and to premultiply (3.25) and (3.27) respectively by $\tilde{B}_0$ and $\tilde{B}_1$ and add them together where $\begin{pmatrix} \tilde{B}_0 & \tilde{B}_1 \end{pmatrix}$ is a highest rank left anihilator of $\begin{pmatrix} B_0 \\ B_1 \end{pmatrix}$. These simplifications allow us to rewrite constraints (3.31) with smaller dimensions.

This type of simplification can be done systematically as described in the following Lemma.

Lemma 3.3 *There exist a full row rank matrix S and a full column rank matrix T such that*

$$J(\beta, v) = \min \sum_{k=0}^{N-1} \nu(k)^T V_\beta \nu(k) \tag{3.35}$$

subject to

$$SET\tilde{\xi}(k+1) = SFT\tilde{\xi}(k) + SG\nu(k) + SHv(k), \tag{3.36}$$

has the same solution as (3.30) subject to constraints (3.31), for all $v(k)$, where

- *SET has full column rank and*
- *$(zSET - SFT)$ has full column rank $\forall z$.*

Proof Let us first put the pencil $\{E, F\}$ in Kronecker form. As shown in [12], there exist orthogonal matrices Q and Z such that

$$Q(zE - F)Z = \begin{pmatrix} zE_\epsilon - F_\epsilon & * & * & * \\ 0 & zE_\infty - F_\infty & * & * \\ 0 & 0 & zE_f - F_f & * \\ 0 & 0 & 0 & zE_\eta - F_\eta \end{pmatrix} \tag{3.37}$$

where the the eigenmodes of the square pencils $\{E_f, F_f\}$ and $\{E_\infty, F_\infty\}$ are respectively the finite and infinite eigenmodes of $\{E, F\}$, $zE_\epsilon - F_\epsilon$ and $zE_\eta - F_\eta$ are respectively full row rank and full column rank, for all z, and E_ϵ and E_η are respectively full row rank and full column rank. Let

$$\begin{pmatrix} \xi_1(j) \\ \xi_2(j) \\ \xi_3(j) \\ \xi_4(j) \end{pmatrix} = Z^T \xi(j). \tag{3.38}$$

Then (3.31) can be expressed as follows

$$\begin{pmatrix} E_\epsilon & * & * & * \\ 0 & E_\infty & * & * \\ 0 & 0 & E_f & * \\ 0 & 0 & 0 & E_\eta \end{pmatrix} \begin{pmatrix} \xi_1(k+1) \\ \xi_2(k+1) \\ \xi_3(k+1) \\ \xi_4(k+1) \end{pmatrix} - \begin{pmatrix} F_\epsilon & * & * & * \\ 0 & F_\infty & * & * \\ 0 & 0 & F_f & * \\ 0 & 0 & 0 & F_\eta \end{pmatrix} \begin{pmatrix} \xi_1(k) \\ \xi_2(k) \\ \xi_3(k) \\ \xi_4(k) \end{pmatrix} = \begin{pmatrix} * \\ * \\ * \\ * \end{pmatrix}. \tag{3.39}$$

It is straightforward to verify that no matter what the values of the first 3 entries of the vector on the right hand side of (3.39) is, there exist sequences ξ_1, ξ_2 and ξ_3 such that (3.39) is satisfied. This means that the top 3 equations in (3.39) do not impose any constraint on ν and v. We can thus take

$$S = \begin{pmatrix} 0 & 0 & 0 & I \end{pmatrix} Q, \quad T = Z \begin{pmatrix} 0 & 0 & 0 & I \end{pmatrix}^T, \tag{3.40}$$

and of course $\tilde{\xi} = \xi_4$. ■

The decomposition (3.37) can be done in a numerically robust way [12]. But since we are only interested in the "η" part, we can also use simpler algorithms; see for example the reduction algorithm introduced in [9].

Thanks to Lemma 3.3, we can assume from here on that

$$E \text{ has full column rank} \tag{3.41}$$
$$\begin{pmatrix} zE - F \end{pmatrix} \text{ has full column rank, } \forall z \tag{3.42}$$
$$\begin{pmatrix} zE - F & G \end{pmatrix} \text{ has full row rank, } \forall z \tag{3.43}$$
$$\begin{pmatrix} E & G \end{pmatrix} \text{ has full row rank.} \tag{3.44}$$

Assumption (3.43) follows from (2.3), and (3.44) the full rankedness of N_i's.

To construct the solution of problem (3.30) subject to constraints (3.31), we use the method of dynamic programming. Let $J_i(\xi(i))$ denote the past cost function:

$$J_i(\xi(i)) = \min \sum_{k=0}^{i-1} \nu(k)^T V_\beta \nu(k) \tag{3.45}$$

subject to

$$E\xi(k+1) = F\xi(k) + G\nu(k) + Hv(k), \quad k = 0, \cdots, i-1. \tag{3.46}$$

Note that J_i depends on given constants v and β but for simplicity of the notations, we do not explicitly express this dependence. Clearly,

$$J(\beta, v) = \min_{\xi(N)} J_N(\xi(N)). \tag{3.47}$$

The J_i's can be constructed recursively as follows:

Lemma 3.4

$$J_i(\xi(i)) = \left(s(i) - \begin{pmatrix} 0 \\ F \end{pmatrix} \xi(i)\right)^T \Gamma_\beta(i) \left(s(i) - \begin{pmatrix} 0 \\ F \end{pmatrix} \xi(i)\right) + \sum_{j=0}^{i-1} s(j)^T \Gamma_\beta(j) s(j) \tag{3.48}$$

where

$$s(j+1) = \begin{pmatrix} \begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(j) s(j) \\ Hv_j \end{pmatrix}, \quad s(0) = 0, \tag{3.49}$$

$$\Gamma_\beta(j+1) = \begin{pmatrix} -\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(j) \begin{pmatrix} 0 \\ E \end{pmatrix} & F^T \\ F & GV_\beta^{-1}G^T \end{pmatrix}^{-1}, \quad \Gamma_\beta(0) = 0. \tag{3.50}$$

Proof The proof is by induction. Clearly $J_0(\xi(0))$ is zero. Now suppose that (3.48) holds for $i = k$ and let us show that it also holds for $i = k+1$. The dynamic programming equation is

$$J_{k+1}(\xi(k+1)) = \min_{\xi(k), \nu(k)} J_k(\xi(k)) + \nu(k)^T V_\beta \nu(k) \tag{3.51}$$

subject to

$$E\xi(k+1) = F\xi(k) + G\nu(k) + Hv(k). \tag{3.52}$$

Introducing the Lagrange multiplier vector λ, we define the Lagrangian

$$\mathcal{L} = J_k(\xi(k)) + \nu(k)^T V_\beta \nu(k) - \lambda^T (E\xi(k+1) - F\xi(k) - G\nu(k) - Hv(k)). \tag{3.53}$$

Setting the partials of $\mathcal{L}$ with respect to $\xi(k)$ and $\nu(k)$ to zero, we get that the solutions $\xi^*(k)$ and $\nu^*(k)$ satisfy

$$\begin{pmatrix} -\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) \begin{pmatrix} 0 \\ E \end{pmatrix} & F^T \\ F & GV_\beta^{-1}G^T \end{pmatrix} \begin{pmatrix} \xi^*(k) \\ \lambda/2 \end{pmatrix} = \begin{pmatrix} -\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) s(k) \\ E\xi(k+1) - Hv(k) \end{pmatrix} \tag{3.54}$$

and $\nu^*(k) = -V_\beta^{-1} G^T \lambda/2$.

The matrix on the left hand side of (3.54) is invertible. This can be proved by first noting that

$$\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) \begin{pmatrix} 0 \\ E \end{pmatrix} \geq 0 \tag{3.55}$$

because, in case $v = 0$, thanks to our assumption,

$$J_k(\xi(k)) = \xi(k)^T \begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) \begin{pmatrix} 0 \\ E \end{pmatrix} \xi(k) \tag{3.56}$$

which is non-negative for all $\xi(k)$. Now suppose that the matrix on the left hand side of (3.54) is not invertible, i.e.,

$$\begin{pmatrix} -\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) \begin{pmatrix} 0 \\ E \end{pmatrix} & F^T \\ F & GV_\beta^{-1}G^T \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0 \tag{3.57}$$

for some vectors x and y (not both zero). From (3.57) we obtain:

$$-\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) \begin{pmatrix} 0 \\ E \end{pmatrix} x + F^T y = 0 \tag{3.58}$$

$$Fx + GV_\beta^{-1}G^T y = 0 \tag{3.59}$$

which implies that

$$x^T \begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) \begin{pmatrix} 0 \\ E \end{pmatrix} x + y^T GV_\beta^{-1}G^T y = 0 \tag{3.60}$$

Thanks to (3.55) and the positivity of V_β, both terms in (3.60) must be zero. Thus

$$\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) \begin{pmatrix} 0 \\ E \end{pmatrix} x = 0 \tag{3.61}$$

$$G^T y = 0 \tag{3.62}$$

which using (3.58) implies that

$$\begin{pmatrix} F^T \\ G^T \end{pmatrix} y = 0. \tag{3.63}$$

But y is not zero (otherwise from (3.59) and Assumption (3.42), we have $x = 0$) which means that (3.63) contradicts (3.43). Thus the matrix on the left hand side of (3.54) is invertible, and so the matrix

$$\Gamma_\beta(k+1) = \begin{pmatrix} -\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) \begin{pmatrix} 0 \\ E \end{pmatrix} & F^T \\ F & GV_\beta^{-1}G^T \end{pmatrix}^{-1} \tag{3.64}$$

is well defined.

We thus get that the solution to the optimization problem (3.51) is unique and is given by

$$\begin{pmatrix} \xi^*(k) \\ \nu^*(k) \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & -V_\beta^{-1}G^T \end{pmatrix} \Gamma_\beta(k+1) \begin{pmatrix} -\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) s(k) \\ E\xi(k+1) - Hv(k) \end{pmatrix}. \tag{3.65}$$

Thus

$$J_{k+1}(\xi(k+1)) = J_k(\xi^*(k)) + \nu^*(k)^T V_\beta \nu^*(k) \tag{3.66}$$

which after some straightforward algebra and using (3.49), can be expressed as follows

$$J_{k+1}(\xi(k+1)) = \left(s(k+1) - \begin{pmatrix}0\\E\end{pmatrix}\xi(k+1)\right)^T \Gamma_\beta(k+1)\left(s(k+1) - \begin{pmatrix}0\\E\end{pmatrix}\xi(k+1)\right) + \sum_{j=0}^{k} s(j)^T\Gamma_\beta(j)s(j). \quad (3.67)$$

But this is just (3.48) for $i = k+1$. ■

Theorem 3.2 *The auxiliary cost function $J(\beta, v)$ is given by*

$$J(\beta, v) = \sum_{j=1}^{N-1} s(j)^T\Gamma_\beta(j)s(j) + s(N)^T\Phi_\beta s(N) \quad (3.68)$$

where the $s(j)$'s and $\Gamma_\beta(j)$'s are respectively defined in (3.49) and (3.50), and where

$$\Phi_\beta = \Gamma_\beta(N) - \Gamma_\beta(N)\begin{pmatrix}0\\E\end{pmatrix}\left(\begin{pmatrix}0 & E^T\end{pmatrix}\Gamma_\beta(N)\begin{pmatrix}0\\E\end{pmatrix}\right)^\dagger \begin{pmatrix}0 & E^T\end{pmatrix}\Gamma_\beta(N). \quad (3.69)$$

Proof To compute $J(\beta, v)$, we can use (3.47) and (3.48) for $j = 0$. It is straightforward to show that any $\xi^*(N)$ satisfying

$$\begin{pmatrix}0 & E^T\end{pmatrix}\Gamma_\beta(N)\begin{pmatrix}0\\E\end{pmatrix}\xi^*(N) = \begin{pmatrix}0 & E^T\end{pmatrix}\Gamma_\beta(N)s(N) \quad (3.70)$$

is an optimal solution of (3.47). Equation (3.68) is then obtained by placing

$$\xi^*(N) = \left(\begin{pmatrix}0 & E^T\end{pmatrix}\Gamma_\beta(N)\begin{pmatrix}0\\E\end{pmatrix}\right)^\dagger \begin{pmatrix}0 & E^T\end{pmatrix}\Gamma_\beta(N)s(N) \quad (3.71)$$

in (3.48). ■

After the auxiliary cost function $J(\beta, v)$, we consider the construction of $J^*(\beta)$ as defined in (3.20). This is not a quadratic optimization problem, however it is possible to convert it to a related quadratic problem:

$$\tilde{J}(\beta) = \max_v J(\beta, v) - \gamma\|v\|^2. \quad (3.72)$$

$J^*(\beta)$ can then be obtained by noting that for $\gamma \geq J^*(\beta)$, the optimization problem (3.72) is well posed and $\tilde{J}(\beta) = 0$, otherwise, $\tilde{J}(\beta) = \infty$. Thus $J^*(\beta)$ can be obtained by a simple γ-iteration algorithm.

Using (3.49), (3.50) and (3.68), the optimization problem (3.72) can be solved recursively (for each γ). This however is not particularly useful for the asymptotic analysis that we shall undertake later, so here we proceed in a more direct way. In particular, we construct an explicit expression for

$J(\beta, v)$. For that, it suffices to express the $s(j)$'s explicitly in terms of v's. Note that System (3.49) is causal, so $s(j)$ depends only on the past values of v. Thus there exists a matrix $Q_\beta(j)$ such that

$$s(j) = Q_\beta(j) \begin{pmatrix} v(0) \\ \vdots \\ v(j-1) \end{pmatrix}. \tag{3.73}$$

With this and Theorem 3.2, we have all the necessary ingredients to state the main result of this section.

Theorem 3.3 *The auxiliary cost function is given by*

$$J(\beta, v) = v^T \Delta_\beta v \tag{3.74}$$

where

$$\Delta_\beta = \begin{pmatrix} \Delta_\beta(N-1) & 0 \\ 0 & 0 \end{pmatrix} + Q_\beta(N)^T \Phi_\beta Q_\beta(N) \tag{3.75}$$

and where

$$\Delta_\beta(k+1) = \begin{pmatrix} \Delta_\beta(k) & 0 \\ 0 & 0 \end{pmatrix} + Q_\beta(k+1)^T \Gamma_\beta(k) Q_\beta(k+1) \tag{3.76}$$

$$\Delta_\beta(0) = [\ ] \quad (\textit{empty matrix}), \tag{3.77}$$

and

$$Q_\beta(k+1) = \begin{pmatrix} \begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta(k) Q_\beta(k) & 0 \\ 0 & H \end{pmatrix} \tag{3.78}$$

$$Q_\beta(1) = \begin{pmatrix} 0 \\ H \end{pmatrix}. \tag{3.79}$$

It is straightforward to see that

$$J^*(\beta) = \rho(\Delta_\beta) \tag{3.80}$$

where $\rho(.)$ denotes the spectral radius[1]. And thus, in Theorem 3.1,

$$\gamma^* = \max_{0 \le \beta \le 1} \rho(\Delta_\beta) \tag{3.81}$$

and w^* is an eigenvector associated with the largest eigenvalue of Δ_{β^*} (which is γ^*) where β^* is a solution of the optimization problem (3.81).

Even though we have not taken advantage of the particular structure of the Δ_β in computing $J^*(\beta)$, as would have a fully recursive approach. This method can still be more efficient because it allows us to use powerful standard linear algebra routines to find the largest eigenvalue and the associated eigenvector of Δ_β.

[1]Since Δ_β is symmetric non-negative, this is just its largest eigenvalue.

3.3 Example

It would take a lot of space to present any meaningful example. So instead, to simply give an idea of what the solutions look like, we consider two randomly generated models. In our example, u and v are of size one, and y of size 2. x_0, x_1, ν_0, ν_1 are of sizes 4, 2, 3, and 4 respectively. Both models are stable. $N = 60$ and $\sigma = 1$.

Figure 3.2 is a plot of $J^*(\beta)$ versus β. In this case, as it often is with randomly generated examples, this function is concave. This is not always the case!

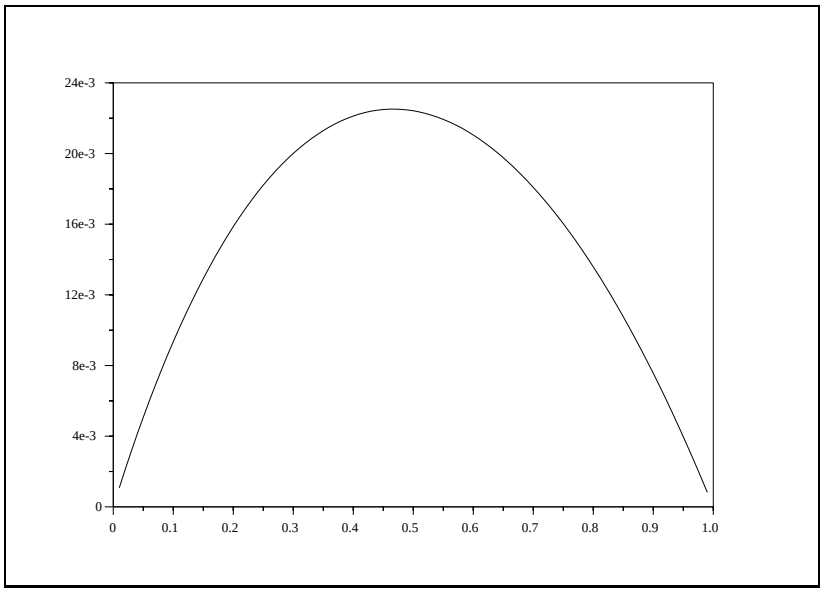

Figure 3.2 : This figure shows $J^*(\beta)$ as a function of β. $\beta^* = .47$ and $\gamma^* = .02251$.

Figure 3.3 illustrates the eigenvector of Δ_{β^*} corresponding to its largest eigenvalue which is equal to γ^*, scaled to give the minimum energy detection signal.

4 Asymptotic behavior

For N large, it is possible to find simple approximate (asymptotically optimal) solutions to the minimum energy detection signal design problem. This allow us to avoid the construction of the Δ_β matrix. We of course continue to assume, as shown previously, that (3.41) through (3.44) hold.

4.1 The Algebraic Riccati Equation

To study the asymptotic behavior of the solution as N goes to infinity, let

$$P_\beta(k) = \begin{pmatrix} 0 & I \end{pmatrix} \Gamma_\beta(k) \begin{pmatrix} 0 \\ I \end{pmatrix}. \tag{4.1}$$

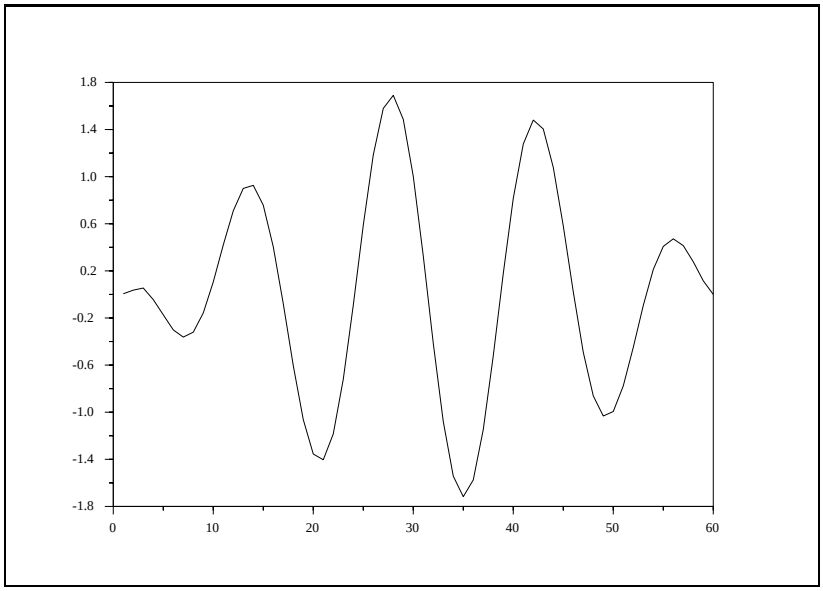

Figure 3.3 : A minimum energy detection signal.

Then, clearly

$$P_\beta(k) = \begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} -E^T P_\beta(k-1) E & F^T \\ F & G V_\beta^{-1} G^T \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ I \end{pmatrix} \tag{4.2}$$

$$P_\beta(0) = 0. \tag{4.3}$$

Theorem 4.1 *$P_\beta(k)$ converges exponentially fast to the unique positive definite solution of the algebraic descriptor Riccati equation*

$$P_\beta = \left(F \left(E^T P_\beta E \right)^{-1} F^T + G V_\beta^{-1} G^T \right)^{-1}. \tag{4.4}$$

Proof The proof has four parts. First we show that $P_\beta(k)$ is increasing. Then we show that it is upper-bounded. This proves that $P_\beta(k)$ converges. Then we show that the limit P_β is positive definite. And finally we show that P_β is the unique solution of the algebraic descriptor Riccati equation (4.4).

Lemma 4.1 *The sequence of $P_\beta(k)$'s satisfies*

$$P_\beta(k+1) \geq P_\beta(k), \quad \forall k \geq 0. \tag{4.5}$$

Proof First consider the optimization problem

$$\mathcal{V}(X, \phi) = \min_{\xi, \nu} \xi^T E^T X E \xi + \nu^T V_\beta \nu \tag{4.6}$$

subject to $F\xi = \phi - G\nu$. Clearly,

$$X_1 \geq X_2 \implies \mathcal{V}(X_1, \phi) \geq \mathcal{V}(X_2, \phi), \ \forall \phi. \tag{4.7}$$

The solution to the optimization problem (4.6) is

$$\mathcal{V}(X,\phi) = \phi^T \begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} -E^T X E & F^T \\ F & G V_\beta^{-1} G^T \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ I \end{pmatrix} \phi. \tag{4.8}$$

Thus by letting $X_1 = P_\beta(k-1)$ and $X_2 = P_\beta(k)$, we get

$$P_\beta(k) \geq P_\beta(k-1) \implies P_\beta(k+1) \geq P_\beta(k) \tag{4.9}$$

but $P_\beta(1) \geq 0 = P_\beta(0)$ so $P_\beta(k)$ is increasing and positive semi-definite. ■

Lemma 4.2 *There exist a positive semi-definite matrix $\hat{P}$ such that*

$$P_\beta(k) \leq \hat{P}, \quad \forall k \geq 0. \tag{4.10}$$

Proof From (3.43), it is easy to see that there exists an invertible matrix

$$Y = \begin{pmatrix} Y_1 & Y_2 \\ Y_3 & Y_4 \end{pmatrix} \tag{4.11}$$

such that

$$\begin{pmatrix} E - zF & zG \end{pmatrix} \begin{pmatrix} Y_1 & Y_2 \\ Y_3 & Y_4 \end{pmatrix} = \begin{pmatrix} zI + EY_1 & EY_2 \end{pmatrix}. \tag{4.12}$$

where $\begin{pmatrix} Y_1 & Y_2 \end{pmatrix}$ is a right inverse of $\begin{pmatrix} -F & G \end{pmatrix}$. Clearly $\begin{pmatrix} zI + EY_1 & EY_2 \end{pmatrix}$ has full row rank, $\forall z \neq 0$, which implies that (EY_1, EY_2), and consequently $(-EY_1, EY_2)$ is a stabilizable pair. Thus there exists a matrix K such that $A = -EY_1 + EY_2K$ is stable (has all of its eigenvalues inside the unit circle). Let

$$\begin{pmatrix} L_1 \\ L_2 \end{pmatrix} = \begin{pmatrix} Y_1 + Y_2K \\ Y_3 + Y_4K \end{pmatrix} \tag{4.13}$$

Then

$$\begin{pmatrix} -F & G \end{pmatrix} \begin{pmatrix} L_1 \\ L_2 \end{pmatrix} = I \tag{4.14}$$

and EL_1 has all of its eigenvalues inside the unit circle.

Now consider the following cost function

$$\hat{J}(z) = \sum_{k=0}^{N-1} \nu(k)^T V_\beta \nu(k) \tag{4.15}$$

subject to

$$E\xi(k+1) = F\xi(k) + G\nu(k), \quad k = 0, \cdots, N-1, \tag{4.16}$$

$$z = \xi(N) \tag{4.17}$$

where we let

$$\nu(k) = L_2 E\xi(k+1). \tag{4.18}$$

This choice of ν yields

$$\xi(k) = -L_1 E\xi(k+1) \tag{4.19}$$

but the nonzero eigenvalues of L_1E are identical to those of EL_1 (which are inside the unit circle), thus recursion (4.19) is stable and $\xi(k)$ converges exponentially to zero. Then $\nu(k)$ also converges to zero thanks to (4.18). This implies that $\hat{J}(z)$ converges as N goes to infinity, for all z, which implies that there exists a positive semi-definite matrix $\hat{Q}$ such that

$$\lim_{N\to\infty} \hat{J}(z) = z^T \hat{Q} z, \ \ \forall z. \tag{4.20}$$

Now consider the same cost function but instead of the particular choice of ν used above, take the ν that minimizes the cost, i.e.,

$$J(\xi(N)) = \min \sum_{k=0}^{N-1} \nu(k)^T V_\beta \nu(k) \tag{4.21}$$

subject to (4.16). This problem is of course the same as problem (3.45) with $v = 0$. The solution is

$$J(\xi(N)) = \xi(N)^T F^T P_\beta(N) F\xi(N), \ \ \forall \xi(N). \tag{4.22}$$

Since the optimal solution is necessarily smaller than or equal to any particular solution, and thanks to the fact that the sequence $P_\beta(k)$ is increasing, we have that

$$E^T P_\beta(k) E \le \hat{Q}, \ \ \forall k, \tag{4.23}$$

which implies (4.10) where

$$\hat{P} = \begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} -\hat{Q} & F^T \\ F & GV_\beta^{-1}G^T \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ I \end{pmatrix}. \tag{4.24}$$

■

So for we have shown that $P_\beta(k)$ is increasing and bounded, which implies that it converges to some P_β. Now we show that P_β is positive definite. Suppose it is not and let X be a matrix such that its columns form a basis for the null space of P_β, and let

$$\begin{pmatrix} S \\ T \end{pmatrix} = \begin{pmatrix} -E^T P_\beta E & F^T \\ F & GV_\beta^{-1}G^T \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ X \end{pmatrix}. \tag{4.25}$$

Note that

$$P_\beta = \begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} -E^T P_\beta E & F^T \\ F & GV_\beta^{-1}G^T \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ I \end{pmatrix} \tag{4.26}$$

and the image of X is in the null space of P_β, so T in (4.25) is zero. Thus, from (4.25) follows that

$$E^T P_\beta E S = 0 \tag{4.27}$$
$$FS = X \tag{4.28}$$

which implies that

$$P_\beta E S = 0 \tag{4.29}$$
$$P_\beta F S = 0 \tag{4.30}$$

and since E and F are full column rank, the columns of ES and FS form two bases for the null space of P_β. Thus there exists a square invertible matrix L such that $ES = FSL$. Let U be the matrix of change of basis that puts L in Jordan form: $L = UJU^{-1}$ where J is in Jordan form. Thus $ESU = FSUJ$, so if we denote the first column of SU by s, we have

$$Es = J_{11} Fs \tag{4.31}$$

where J_{11} is the $(1,1)$ entry of J (because J is upper triangular[2]). S has full column rank (because X has full column rank), so SU has full column rank which implies that s is not zero. But then (4.31) contradicts Assumption (3.42). Thus P_β is positive definite.

Finally, we must show that there is a unique positive definite solution to the algebraic descriptor Riccati equation. Suppose there are two distinct solutions P_1 and P_2, i.e.,

$$P_i = \left(F \left(E^T P_i E \right)^{-1} F^T + G V_\beta^{-1} G^T \right)^{-1}, \quad i = 1, 2. \tag{4.32}$$

By taking the inverse of both sides of (4.32) and subtracting the result for $i = 2$ from the result for $i = 1$, we get

$$P_1^{-1} - P_2^{-1} = F(E^T P_1 E)^{-1} E^T P_1 (P_1^{-1} - P_2^{-1}) (F(E^T P_2 E)^{-1} E^T P_2)^T \tag{4.33}$$

which implies that, for all $k \geq 1$,

$$P_1^{-1} - P_2^{-1} = (F(E^T P_1 E)^{-1} E^T P_1)^k (P_1^{-1} - P_2^{-1}) ((F(E^T P_2 E)^{-1} E^T P_2)^T)^k. \tag{4.34}$$

Clearly if we show that $F(E^T P_i E)^{-1} E^T P_i$, $i = 1, 2$, have all their eigenvalues inside the unit circle, we immediately have that $P_1 = P_2$. This can be shown by noting that thanks to (4.32),

$$P_i^{-1} - (F(E^T P_i E)^{-1} E^T P_i) P_i^{-1} (F(E^T P_i E)^{-1} E^T P_i)^T = G V_\beta^{-1} G^T. \tag{4.35}$$

But (4.35) is a Lyapunov equation and thus it is enough to show that

$$(F(E^T P_i E)^{-1} E^T P_i,\ G), \quad i = 1, 2,$$

are controllable pairs. Suppose this is not the case, i.e., there exists a z and a non zero w such that

$$w^T \begin{pmatrix} zI - F(E^T P_i E)^{-1} E^T P_i & G \end{pmatrix} = 0 \tag{4.36}$$

[2]Note that s is an eigenvector of $\{E, F\}$, in fact each column of SU is either an eigenvector or a generalized eigenvector of $\{E, F\}$.

which implies that

$$w^T F(E^T P_i E)^{-1} E^T P_i = z w^T \tag{4.37}$$

$$w^T G = 0. \tag{4.38}$$

Multiplying (4.37) on the right by E, and using (4.38), we obtain

$$w^T \begin{pmatrix} zE - F & G \end{pmatrix} = 0 \tag{4.39}$$

which is clearly a contradiction (see Assumption (3.43)). Thus, both matrices $F(E^T P_1 E)^{-1} E^T P_1$ and $F(E^T P_2 E)^{-1} E^T P_2$ have all their eigenvalues inside the unit circle. ■

The solution to the algebraic descriptor Riccati equation (4.4) can be constructed using the matrix pencil

$$\Psi = \left\{ \begin{pmatrix} F & G V_\beta^{-1} G^T \\ 0 & E^T \end{pmatrix}, \begin{pmatrix} E & 0 \\ 0 & F^T \end{pmatrix} \right\}. \tag{4.40}$$

Theorem 4.2 *The matrix pencil Ψ is regular, has no eigenmode on the unit circle and if the columns of $\begin{pmatrix} \Gamma_1 \\ \Gamma_2 \end{pmatrix}$ form a basis for the stable eigenspace of Ψ, i.e.,*

$$\begin{pmatrix} F & G V_\beta^{-1} G^T \\ 0 & E^T \end{pmatrix} \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \end{pmatrix} \mathcal{J} = \begin{pmatrix} E & 0 \\ 0 & F^T \end{pmatrix} \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \end{pmatrix} \tag{4.41}$$

where eigenvalues of $\mathcal{J}$ are inside the unit circle, then,

$$P_\beta = (F \Gamma_1 \Gamma_2^{-1} + G V_\beta^{-1} G^T)^{-1} \tag{4.42}$$

is the unique positive definite solution of the algebraic descriptor Riccati equation (4.4).

Proof Let

$$\Psi(z) = z \begin{pmatrix} F & G V_\beta^{-1} G^T \\ 0 & E^T \end{pmatrix} - \begin{pmatrix} E & 0 \\ 0 & F^T \end{pmatrix} = \begin{pmatrix} zF - E & z G V_\beta^{-1} G^T \\ 0 & zE^T - F^T \end{pmatrix}. \tag{4.43}$$

Suppose z is on the unit circle and let z^* denote the complex conjugate of z. Note that $z^* = 1/z$. To show that z is not an eigenmode of Ψ, we must show that $\Psi(z)$ is invertible, or equivalently that

$$\begin{pmatrix} zF - E & G V_\beta^{-1} G^T \\ 0 & z^* F^T - E^T \end{pmatrix}$$

is invertible. Suppose this is not the case, i.e., there exist x and y, not both zero, such that

$$\begin{pmatrix} zF - E & G V_\beta^{-1} G^T \\ 0 & z^* F^T - E^T \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0. \tag{4.44}$$

But this implies, after premultiplication of the first equation by the complex conjugate transpose of y, that

$$\begin{pmatrix} G^T \\ z^* F^T - E^T \end{pmatrix} y = 0 \tag{4.45}$$

Thanks to Assumption (3.43), (4.45) implies that $y = 0$, which in turn implies that

$$(zF - E)\, x = 0 \tag{4.46}$$

which implies that $x = 0$, thank to Assumption (3.42). But this is a contradiction, so Ψ has no eigenmode on the unit circle.

Let $p(z)$ denote the determinant of $\Psi(z)$, and μ the degree of $p(z)$. Thanks to the identity

$$\left[\begin{pmatrix} 0 & -I \\ z^{-1}I & 0 \end{pmatrix} \Psi(z)\right]^T = \begin{pmatrix} 0 & -I \\ zI & 0 \end{pmatrix} \Psi(z^{-1}), \tag{4.47}$$

by taking the determinant of both sides, we get $z^{-m}p(z) = z^m p(z^{-1})$ where m equals the number of rows of E. So, since p does not have any roots on the unit circle, $\mu = m$ and consequently Γ_2 is square.

From (4.41), we get

$$F\Gamma_1\mathcal{J} + GV_\beta^{-1}G^T\Gamma_2\mathcal{J} = E\Gamma_1 \tag{4.48}$$

$$E^T\Gamma_2\mathcal{J} = F^T\Gamma_2 \tag{4.49}$$

which implies that

$$\Gamma_2^T F\Gamma_1 = \mathcal{J}^T\Gamma_2^T F\Gamma_1\mathcal{J} + \mathcal{J}^T\Gamma_2^T GV_\beta^{-1}G^T\Gamma_2\mathcal{J} \tag{4.50}$$

which is a Lyapunov equation and since $\mathcal{J}$ has all its eigenvalues inside the unit circle,

$$W = \Gamma_2^T F\Gamma_1 \tag{4.51}$$

is symmetric positive semi-definite.

Lemma 4.3 *The matrix Γ_2 is invertible.*

Proof Suppose $\Gamma_2 w = 0$ which implies that $Ww = 0$. Thanks to (4.50), we get that

$$W\mathcal{J}w = 0 \tag{4.52}$$

$$G^T\Gamma_2\mathcal{J}w = 0. \tag{4.53}$$

But from (4.49) we get $E^T\Gamma_2\mathcal{J}w = 0$ which thanks to (4.53) and (3.44) implies that $\Gamma_2\mathcal{J}w = 0$. Thus $\ker(\Gamma_2)$ is $\mathcal{J}$-invariant. This implies that there exists at least one eigenvector of $\mathcal{J}$ in $\ker(\Gamma_2)$, i.e., there exist a non-zero vector v and a scalar λ such that $\Gamma_2 v = 0$ and $\mathcal{J}v = \lambda v$. So by multiplying (4.48) on the right by v we obtain

$$(\lambda F - E)\Gamma_1 v = 0, \tag{4.54}$$

which thanks to (3.43) and (3.44) implies that $\Gamma_1 v = 0$. But this is a contradiction because $\begin{pmatrix}\Gamma_1 \\ \Gamma_2\end{pmatrix}$ has full column rank. Thus Γ_2 is invertible. ■

Lemma 4.4 *The following always holds*

$$\ker(F^T\Gamma_2) = \ker(\Gamma_1). \tag{4.55}$$

Proof Since F has full column rank and Γ_2 is invertible, $\ker(W) = \ker(\Gamma_1)$ and since W is symmetric

$$\ker(F^T\Gamma_2) \subset \ker(\Gamma_1). \tag{4.56}$$

Now we show that

$$\ker(\mathcal{J}) = \ker(F^T\Gamma_2). \tag{4.57}$$

From (4.49) it follows that $\ker(\mathcal{J}) \subset \ker(F^T\Gamma_2)$. Let w be any vector such that $F^T\Gamma_2 w = 0$, this implies, thanks to (4.50), that

$$G^T\Gamma_2\mathcal{J}w = 0 \tag{4.58}$$

and thanks to (4.49), that

$$E^T\Gamma_2\mathcal{J}w = 0. \tag{4.59}$$

But (4.58) and (4.59), because of Assumption (3.44), imply that $\Gamma_2\mathcal{J}w$ and consequently $\mathcal{J}w$ is zero. Thus $\ker(F^T\Gamma_2) \subset \ker(\mathcal{J})$. This proves (4.57).

Now we show that

$$\ker(\mathcal{J}) = \ker(\Gamma_1). \tag{4.60}$$

From (4.48) and the full rankedness of E, it is easy to see that $\ker(\mathcal{J}) \subset \ker(\Gamma_1)$.

Let w be any vector satisfying $\Gamma_1 w = 0$, then by premultiplying by w^T and postmultiplying by w (4.50), we obtain

$$\begin{aligned} \Gamma_1\mathcal{J}w &= 0 & (4.61)\\ G^T\Gamma_2\mathcal{J}w &= 0. & (4.62) \end{aligned}$$

From (4.61) we get that $\ker(\Gamma_1)$ is $\mathcal{J}$-invariant. Let v be any eigenvector of $\mathcal{J}$ satisfying $\Gamma_1 v = 0$, and λ the associated eigenvalue, i.e., $\mathcal{J}v = \lambda v$. Then,

$$\begin{aligned} \lambda E^T\Gamma_2 v &= F^T\Gamma_2 v = 0 & (4.63)\\ \lambda G^T\Gamma_2 v &= 0. & (4.64) \end{aligned}$$

Thus λ is necessarily zero. So the restriction of $\mathcal{J}$ to $\ker\Gamma_1$ is nilpotent. We denote it by $\mathcal{N}$.

Now suppose $\ker(\Gamma_1)$ is not a subset of $\ker(\mathcal{J})$, i.e., there exists a vector w such that $\Gamma_1 w = 0$ but $\mathcal{J}w \neq 0$. This clearly implies that $\mathcal{N} \neq 0$, i.e., the nilpotent matrix $\mathcal{N}$ has non trivial Jordan blocks which in turn implies that there exists a vector $v \in \ker(\Gamma_1)$ such that

$$\begin{aligned} \mathcal{J}v &\neq 0 & (4.65)\\ \mathcal{J}^2 v &= 0. & (4.66) \end{aligned}$$

But then from (4.50) and (4.48) follows that

$$\begin{aligned} F^T\mathcal{J}v &= 0 & (4.67)\\ G^T\mathcal{J}v &= 0 & (4.68) \end{aligned}$$

which because of Assumption (3.43), imply that $\mathcal{J}v = 0$. But this is a contradiction. This shows (4.60). Finally, (4.55) follows from (4.57) and (4.60). ■

Now we can show that

$$\mathcal{H} = F\Gamma_1\Gamma_2^{-1} + GV_\beta^{-1}G^T \tag{4.69}$$

is invertible. Note that $F\Gamma_1\Gamma_2^{-1} = (\Gamma_2^{-1})^T W \Gamma_2^{-1}$ is positive semi-definite. So if $\mathcal{H}w = 0$, then

$$W\Gamma_2^{-1}w = 0 \tag{4.70}$$

$$G^T w = 0. \tag{4.71}$$

But thanks to (4.55) and full rankedness of F, $\ker(W) = \ker(F^T\Gamma_2)$. Thus (4.70) and (4.71) imply that $w^T \begin{pmatrix} F & G \end{pmatrix} = 0$, which implies that $w = 0$. Thus $\mathcal{H}$ is invertible and positive-definite. From (4.48) follows that $\Gamma_2\mathcal{J} = \mathcal{H}^{-1}E\Gamma_1$ and from (4.49), that

$$E^T\mathcal{H}^{-1}E\Gamma_1\Gamma_2^{-1} = F^T. \tag{4.72}$$

But E has full column rank, so that $E^T\mathcal{H}^{-1}E$ is invertible. Thus from (4.72) we obtain $F\Gamma_1\Gamma_2^{-1} = F(E^T\mathcal{H}^{-1}E)^{-1}F^T$. But then thanks to (4.69), we obtain

$$\mathcal{H} = GV_\beta^{-1}G^T + F(E^T\mathcal{H}^{-1}E)^{-1}F^T. \tag{4.73}$$

By letting

$$P_\beta = \mathcal{H}^{-1}, \tag{4.74}$$

we obtain the algebraic descriptor Riccati equation (4.4). Noting that (4.74) is equivalent to (4.42) and Theorem 4.2 is proved. ■

4.2 Construction of detection signal for large N

For the construction of a detection signal for large N, we need to study the asymptotic behavior of the auxiliary cost function $J(\beta, v)$ as N goes to infinity. From here on, we denote the detection signal v_N and the the auxiliary cost function $J_N(\beta, v)$ to show their dependence on N (so far we had considered N to be fixed).

Note that the convergence of $P_\beta(k)$ to P_β clearly implies the convergence of $\Gamma_\beta(k)$ to

$$\Gamma_\beta = \begin{pmatrix} -E^T P_\beta E & F^T \\ F & GV_\beta^{-1}G^T \end{pmatrix}^{-1}. \tag{4.75}$$

This convergence is exponential and allows us, as will be shown later, for large N, to consider the stationary version of recursion (3.49) for the construction of detection signals. To study the asymptotic behavior of the optimal detection signals, we need to introduce the concept of "asymptotically optimal":

Definition 4.1 *Consider the optimization problem*

$$\max_x f_N(x). \tag{4.76}$$

Then $\overline{x}_N$ is called an asymptotically optimal solution of (4.76) if

$$\lim_{N\to\infty} \{f_N(\overline{x}_N) - \max_x f_N(x)\} = 0 \tag{4.77}$$

.

Let us now consider a new auxiliary cost function

$$\tilde{J}_N(\beta, v_N) = \sum_{j=1}^{N} s(j)^T \Gamma_\beta s(j) - s(N)^T \Gamma_\beta \begin{pmatrix} 0 \\ E \end{pmatrix} \left(E^T P_\beta E\right)^{-1} \begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta s(N) \tag{4.78}$$

where

$$s(j+1) = \begin{pmatrix} 0 & E^T \\ 0 & 0 \end{pmatrix} \Gamma_\beta s(j) + \begin{pmatrix} 0 \\ H \end{pmatrix} v_N(j), \quad s(0) = 0 \tag{4.79}$$

with $0 \leq j \leq N-1$. Note that (4.79) and (4.78) are the stationary versions of (3.49) and (3.68), so we call $\tilde{J}_N$ the stationary auxiliary cost function. It is straightforward to verify that

$$\tilde{J}_N(\beta, v_N) = \min \sum_{k=0}^{N-1} \|\nu(k)\|^2 + \xi(0)^T F^T P_\beta F \xi(0) \tag{4.80}$$

subject to

$$E\xi(k+1) = F\xi_(k) + G\nu(k) + Hv_N(k). \tag{4.81}$$

To see this, simply note that this optimization problem can be solved exactly the same way we solved problem (3.30), the only difference is that because of the initial cost on $\xi(0)$, the cost to go function in the forward dynamic programming approach is not initialized to zero. In fact this particular choice of initial cost results in constant $\Gamma_\beta(i)$'s $(= \Gamma_\beta)$.

This problem is to be compared with (3.30) and (3.31). Clearly $\tilde{J}_N(\beta, v_N) \geq J_N(\beta, v_N)$. The following Lemma which we will prove later in this section, shows that we can study the stationary auxiliary cost function instead of the auxiliary cost function.

Lemma 4.5 *Any asymptotically optimal solution of the stationary auxiliary cost function is an asymptotically optimal solution of the auxiliary cost function, i.e.,*

$$\lim_{N\to\infty} \left(\max_{\substack{v_N \neq 0 \\ \textit{subject to (4.79)}}} \frac{\tilde{J}_N(\beta, v_N)}{\|v_N\|^2} - \max_{\substack{v_N \neq 0 \\ \textit{subject to (3.49)}}} \frac{J_N(\beta, v_N)}{\|v_N\|^2} \right) = 0. \tag{4.82}$$

So, we consider the optimization problem

$$\gamma_N(\beta) = \max_{\substack{v_N \neq 0 \\ \text{subject to (4.79)}}} \frac{\tilde{J}_N(\beta, v_N)}{\|v_N\|^2}. \tag{4.83}$$

Theorem 4.3 *The optimization problem (4.83) is equivalent to*

$$\gamma_N(\beta) = \max_{v_N \neq 0} \frac{\|\zeta\|^2}{\|v_N\|^2} \tag{4.84}$$

where $\mathcal{S}_\beta$ is the stable linear system

$$\mathcal{S}_\beta : \begin{cases} t(j+1) &= F(E^T P_\beta E)^{-1} E^T P_\beta t(j) + Hv_N(j) \\ \zeta(j) &= W_\beta t(j) \end{cases} \tag{4.85}$$

with $t(0) = 0$ and where W_β is any matrix satisfying

$$W_\beta^T W_\beta = P_\beta - P_\beta E(E^T P_\beta E)^{-1} E^T P_\beta. \tag{4.86}$$

Proof Let

$$t(k) = \begin{pmatrix} F(E^T P_\beta E)^{-1} & I \end{pmatrix} s(k), \quad 0 \le k \le N. \tag{4.87}$$

Then, using (4.79), we get

$$t(j+1) = \begin{pmatrix} F(E^T P_\beta E)^{-1} & I \end{pmatrix} \left(\begin{pmatrix} 0 & E^T \\ 0 & 0 \end{pmatrix} \Gamma_\beta s(j) + \begin{pmatrix} 0 \\ H \end{pmatrix} v_N(j) \right) \tag{4.88}$$

which implies the first equation in (4.85). It is also easy to show that

$$s(k+1) = \begin{pmatrix} E^T P_\beta t(k) \\ H v_N(k) \end{pmatrix}, \quad 0 \le k \le N-1. \tag{4.89}$$

From (4.78) and (4.89), after some algebra, we get

$$\tilde{J}_N(\beta, v_N) = \sum_{j=0}^{N-1} t(j+1)^T P_\beta t(j+1) - t(j)^T P_\beta E (E^T P_\beta E)^{-1} E^T P_\beta t(j) - s(N)^T \Gamma_\beta \begin{pmatrix} 0 \\ E \end{pmatrix} (E^T P_\beta E)^{-1} \begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta s(N) \tag{4.90}$$

which, since $t(0) = 0$ and $\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta s(N) = E^T P_\beta t(N)$, implies that

$$\tilde{J}_N(\beta, v_N) = \sum_{j=1}^{N} t(j)^T (P_\beta - P_\beta E (E^T P_\beta E)^{-1} E^T P_\beta) t(j) = \sum_{j=0}^{N-1} \|\zeta(i)\|^2. \tag{4.91}$$

What remains to be shown is that $\mathcal{S}_\beta$ is stable, i.e., that the eigenvalues of $F(E^T P_\beta E)^{-1} E^T P_\beta$ are inside the unit circle. But this matrix is just Λ_2 in the proof of Theorem 4.1 and it is shown there to have all its eigenvalues inside the unit circle. ■

Clearly as N goes to infinity, $\gamma_N(\beta)$ converges to

$$\gamma_\infty(\beta) = \|\mathcal{S}_\beta\|_\infty^2 = \max_\omega \bar{\sigma}(S_\beta(\exp(\sqrt{-1}\omega))^2 \tag{4.92}$$

where $\bar{\sigma}$ denotes the largest singular value and where $S_\beta(z)$ denotes the transfer function associated with System $\mathcal{S}_\beta$:

$$S_\beta(z) = W_\beta (zI - F(E^T P_\beta E)^{-1} E^T P_\beta)^{-1} H. \tag{4.93}$$

The H_∞ norm of a discrete system can be computed by transforming it into a continuous system by a bilinear transformation, which preserves the H_∞ norm, or directly as described in [3]. Thus, $\gamma_\infty(\beta)$ and the critical frequency $\omega = \omega(\beta)$ achieving the maximum in (4.92) can be computed using standard algorithms. Note that $\omega(\beta)$ is a frequency at which, for $N = \infty$, the system $\mathcal{S}_\beta$ has highest gain. This is equal to $\sqrt{\gamma_\infty(\beta)}$.

In case of scalar v_N, an asymptotically optimal solution of (4.83), is

$$v_N(i) = \sin(\frac{\pi i}{N}) \cos(\omega(\beta) i), \quad i = 0, ..., N-1. \tag{4.94}$$

The choice of the envelope (in this case $\sin(\frac{\pi i}{N})$) is of course not unique. We can consider for example $1/N$ for $i = 0, ..., N-1$, or any other function as long as when N goes to infinity, the spectrum of v_N converges to a delta function at critical frequency $\omega(\beta)$.

To show asymptotic optimality of (4.94), we simply have to make sure that as N goes to infinity, v_N converges to a pure sinusoid with critical frequency $\omega(\beta)$. But this follows from

$$v_N(i) = \frac{1}{2}(\sin((\omega(\beta) + \frac{\pi}{N})i) - \sin((\omega(\beta) - \frac{\pi}{N})i)). \tag{4.95}$$

Thus

$$\lim_{N\to\infty} \frac{\tilde{J}_N(\beta, v_N)}{\|v_N\|^2} = \gamma_\infty(\beta) \tag{4.96}$$

proving the asymptotic optimality of v_N.

Note that the norm of v_N can be computed using

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n-1} \cos^2(ai+b)}{n} = \begin{cases} 1 & \text{if } a \text{ and } b \text{ are multiples of } \pi, \\ 1/2 & \text{if } a \text{ is not a multiple of } \pi. \end{cases} \tag{4.97}$$

and the fact that

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n-1} \sin^2(\frac{\pi i}{n})}{n} = 1/2. \tag{4.98}$$

We obtain (clearly if $\omega(\beta) = 0, \pi$, then ϕ is a multiple of π) in particular that $\|v_N\| = \frac{1}{4}$ if $\omega(\beta) \neq 0, \pi$ and $\frac{1}{2}$ otherwise.

It is straightforward to generalize this result to the non-scalar case and construct a complete solution to the optimization problem associated with the stationary auxiliary problem, which is also a solution to our original problem thanks to Lemma 4.5.

Theorem 4.4 *Consider the problem of minimum energy proper detection signal design over* $[0, N]$. *Then, an asymptotically optimal solution* v_N *is given by*

$$v_N(i) = \alpha\sqrt{\frac{\sigma}{N\gamma_\infty(\beta^*)}}\sin(\frac{\pi i}{N}) \begin{pmatrix} p_1\cos(\omega(\beta^*)i + \phi_1) \\ p_2\cos(\omega(\beta^*)i + \phi_2) \\ \vdots \\ p_n\cos(\omega(\beta^*)i + \phi_n) \end{pmatrix} \tag{4.99}$$

where $(p_1\exp(\sqrt{-1}\phi_1) \quad p_2\exp(\sqrt{-1}\phi_2) \quad .. \quad p_n\exp(\sqrt{-1}\phi_n))^T$ *is a normalized eigenvector associated with* $\gamma_\infty(\beta^*)$, *the largest eigenvalue of the matrix* $T_{\beta^*}(z(\beta^*))$ *and where* β^* *satisfies*

$$\max_{0<\beta<1} \gamma_\infty(\beta) = \gamma_\infty(\beta^*). \tag{4.100}$$

The scalar coefficient α *is equal to* 2 *if* $\omega(\beta^*) \neq 0, \pi$; *it is equal to* $\sqrt{2}$ *otherwise.*

Of course this result uses Lemma 4.5 which we have not proved yet.

Proof of lemma 4.5 To prove this Lemma, what we need to show is that in the optimization problem (4.80), for any asymptotically optimal v_N, the corresponding optimal $\xi(0)$ converges to zero as N goes to infinity. Let us consider the asymptotically optimal v_N found in (4.99). Using the fact that this v_N is a stationary (in particular sinusoidal) function of amplitude in the order of $1/\sqrt{N}$ and that the recursion (4.79) is stable, we get that s is a stationary process of amplitude in the order of $1/\sqrt{N}$, and thus $s(N)$ is in the order of $1/\sqrt{N}$. But then

$$\xi^*(N) = \left(\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta \begin{pmatrix} 0 \\ E \end{pmatrix} \right)^\dagger \begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta s(N) \tag{4.101}$$

where $\xi^*(N)$ is the optimal $\xi(N)$. This is just a straightforward extension of Theorem 3.2 to the stationary case. So the optimal $\xi(N)$ converges to zero as N goes to infinity. Once we have $\xi^*(N)$, the other optimal ξ's can be obtained by

$$\xi^*(k) = \begin{pmatrix} I & 0 \end{pmatrix} \Gamma_\beta(k+1) \begin{pmatrix} -\begin{pmatrix} 0 & E^T \end{pmatrix} \Gamma_\beta s(k) \\ E\xi^*(k+1) - Hv_N(k) \end{pmatrix} \tag{4.102}$$

(this is just an extension of (3.65) to the stationary case). But it is straightforward to show that (4.102) is a stable recursion and moreover both v_N and s are stationary processes of amplitude in the order of $1/\sqrt{N}$. Thus $\xi^*(0)$ converges to zero as N goes to infinity. ■

The conclusion of this section is that approximate minimum energy proper detection signal design, for large N, reduces to solving the following simple scalar nonlinear optimization problem

$$\gamma^* = \gamma_\infty(\beta^*) = \max_{0 \le \beta \le 1} \|S(z)\|_\infty. \tag{4.103}$$

An asymptotically optimal solution is then given by (4.99). Note that $\sqrt{\gamma^*}$ is the approximation of the separability index (converging to the real one as N goes to infinity). The above computations can easily be implemented using programs such as Scilab and Matlab.

4.3 Example

Going back to the random example we treated in Section 3.3, Figure 4.1 illustrates the resulting $\gamma_\infty(\beta)$; note the resemblance with Figure 3.2 . The optimal value of γ_∞ is $\gamma^* = .02266$.

Figure 4.2 shows the corresponding detection signal obtained using (4.99). Note that the real separability index for this example is .150033. The approximation of the separability index, obtained by considering the infinite horizon problem, is .150532 which, as expected, is larger than the real one. Finally the effective separability index obtained by using the approximate solution v_N obtained in (4.99), is

$$\sqrt{\frac{1}{\|v_N\|^2} \max_\beta \tilde{J}_N(\beta, v_N)} = \sqrt{\frac{1}{\|v_N\|^2} \max_\beta v_N^T \Delta_\beta v_N} = 0.149666. \tag{4.104}$$

That amounts to an error of less than .3%. Thus, for all practical purposes, the approximate solution can be used instead of the optimal solution. This is true because $N = 60$. For small N, the situation is different. For example for $N = 10$, the real separability index is 0.13153. The use of the approximate solution v_N obtained in (4.99) gives an effective separability index of 0.07603. This corresponds to more than 40% error.

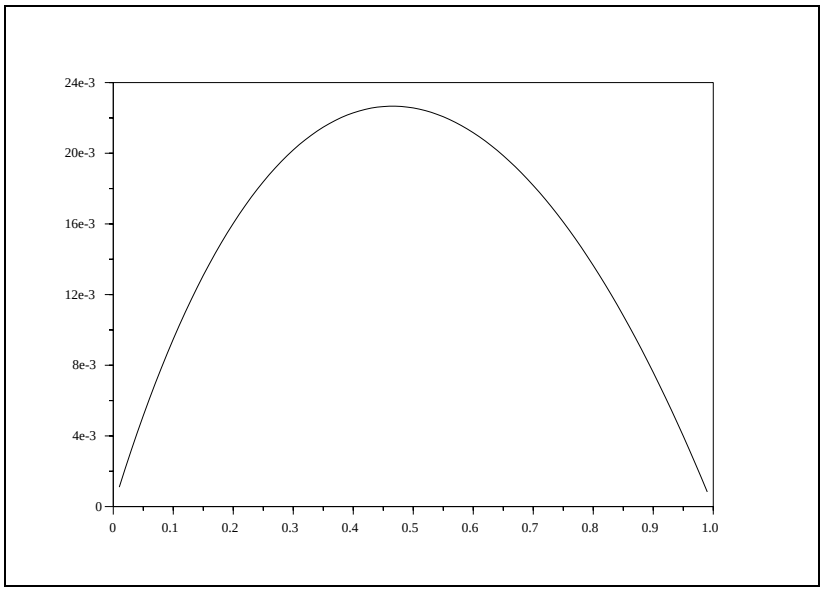

Figure 4.1 : This figure shows $\gamma_\infty(\beta)$ as a function of β. $\beta^* = .47$ and $\gamma^* = .02266$.

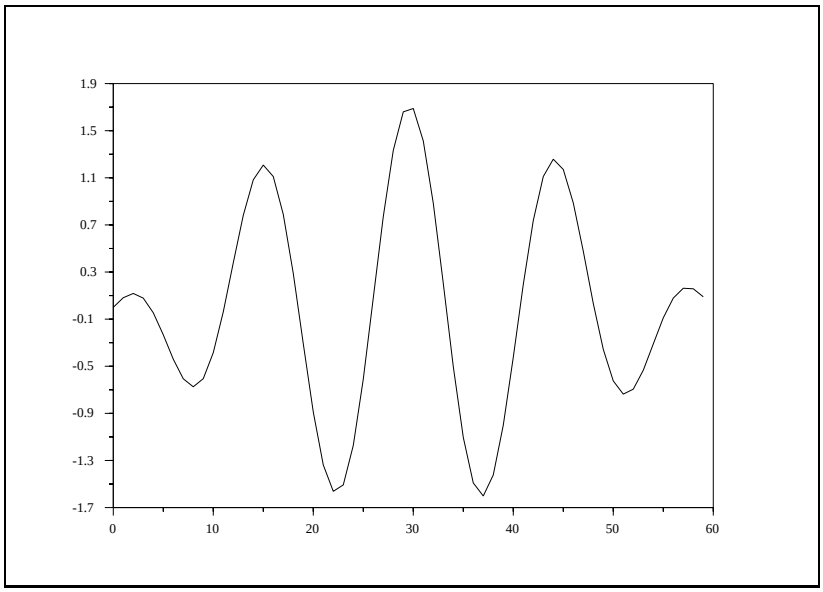

Figure 4.2 : The approximate optimal detection signal, to be compared with the optimal detection signal of Figure 3.3 .

4.4 Discussion

In a real application of the method presented here, one has to have an estimate of σ. It clearly makes sense to assume that σ is proportional to N. In which case, without any loss of generality we can let

$$\sigma = N. \tag{4.105}$$

To see the trade-off between the length of the detection horizon N, and the average power

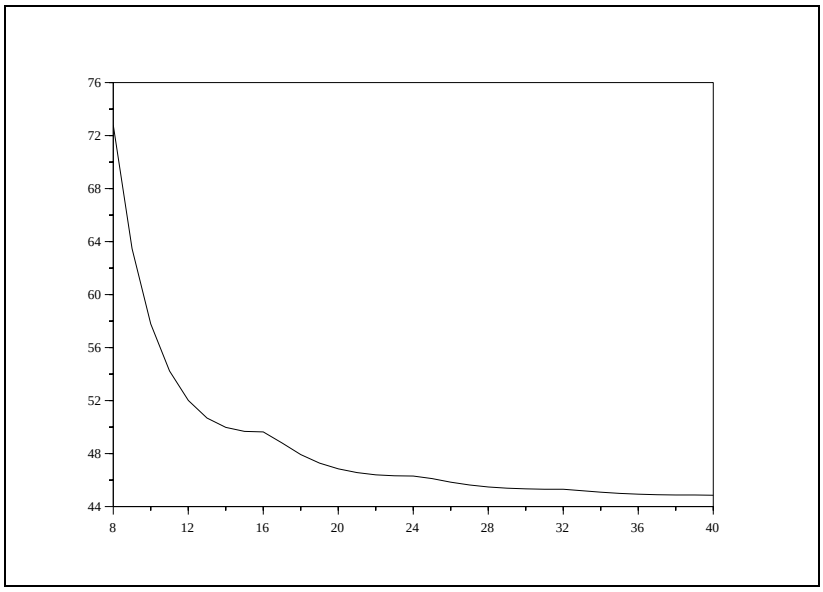

Figure 4.3 : The minimum average power needed for perfect detection as a function of the length of the detection horizon N. Note that as N goes to infinity, this function converges to $1/.02266 = 44.12$.

$\|v_N\|^2/N$ of the optimum detection signal, we have computed the average power for different values of N for our randomly selected example; see Figure 4.3 . As expected, the result is a decreasing function of N. Larger N implies that we can better shape the detection signal and have a higher separability index. The price to pay is of course longer detection horizon.

5 On-line detection filter

Once the detection signal is constructed, we need to design an on-line detection filter to decide, based on measurements u and y, whether the system is in normal mode or in failed mode. If the detection signal is proper, then a correct decision can be made in every situation. Figure 5 illustrates this point. In a perfect world where the behavior of the real process is perfectly captured by our model, the measurements (u, y) would fall either in $\mathcal{A}^0(v)$ or $\mathcal{A}^1(v)$, and not elsewhere. In that case, it suffices to verify for example that (u, y) is in $\mathcal{A}^0(v)$. This can be done by the following test

$$J^0 < \sigma \tag{5.1}$$

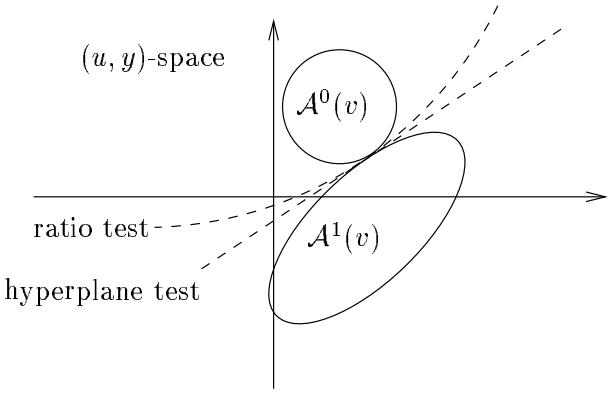

Figure 5.1 : $\mathcal{A}^0(v)$ and $\mathcal{A}^1(v)$ are generalized ellipsoid open sets in the space of measurements (u, y). If v is a minimum energy proper detection signal, then the two sets don't intersect but are at zero distance of each other.

where

$$J^i = \min_{\substack{\nu_i,\, x_i \\ \text{subject to (2.1)-(2.2)}}} \sum_{k=0}^{N-1} \|\nu_i(k)\|^2, \quad i = 0, 1. \tag{5.2}$$

If (5.2) holds, (u, y) is in $\mathcal{A}^0(v)$, and we decide that no failure has occurred, otherwise we decide that a failure has occurred. Thus this test divides the (u, y) space in two parts: $\mathcal{A}^0(v)$ and its complement.

But in the real world, measurements can be outside both sets[3]. If a measurement falls right off the top of $\mathcal{A}^0(v)$ (see Figure 5), the test (5.1) would decide on a failure even thought this measurement is much "closer to" $\mathcal{A}^0(v)$ than it is to $\mathcal{A}^1(v)$. This shows that we need a more "reasonable" way of dividing the (u, y) space. Two such tests are: the hyperplane test, and the ratio test.

5.1 Hyperplane test

The sets $\mathcal{A}^0(v)$ and $\mathcal{A}^1(v)$ are generalized ellipsoids[4]. Thus they can be separated by a hyperplane because any two disjoint convex sets can be separated by a hyperplane.

The advantage of dividing the (u, y) space using this hyperplane is that the corresponding decision test is very simple:

$$\begin{pmatrix} u(0)^T & y(0)^T & u(1)^T & \cdots & u(N-1)^T & y(N-1)^T \end{pmatrix} r < \delta \tag{5.3}$$

for a vector r and a scalar δ. r and δ are the parameters of the hyperplane and can be computed off-line. So the detection filter needs only store the precomputed values of r and δ.

Note however that r and δ depend on the detection signal v. Since such a detection filter is necessarily customized for a specific detection signal, changing the detection signal implies redesigning the filter.

[3] This should not happen very often, otherwise the models should be adjusted

[4] The set X is a generalized ellipsoid if $X = \{x | (x - x_0)^T Q (x - x_0) < 1\}$ for some x_0 and a positive semi-definite matrix Q.

5.2 Ratio test

The ratio test is

$$J^0/J^1 < 1 \tag{5.4}$$

where the J^i's are defined in (5.2). This test, which is reminiscent of the log-likelihood ratio test in the stochastic formulation, can be implemented recursively. Note that

$$J^i = \min_{\nu_i, x_i} \sum_{k=0}^{N-1} \|\nu_i(k)\| \tag{5.5}$$

subject to

$$E_i x_i(k+1) = F_i x_i(k) + G_i \nu_i(k) + H_i \begin{pmatrix} v(k) \\ u(k) \\ y(k) \end{pmatrix}, \tag{5.6}$$

where

$$E_i = \begin{pmatrix} I \\ 0 \end{pmatrix}, \quad F_i = \begin{pmatrix} A_i \\ C_i \end{pmatrix}, \tag{5.7}$$

$$G_i = \begin{pmatrix} M_i \\ N_i \end{pmatrix}, \quad H_i = \begin{pmatrix} D_i & B_i & 0 \\ 0 & 0 & I \end{pmatrix}. \tag{5.8}$$

But this problem is similar to problem (3.24). All the assumptions are verified and we can construct the solution as we did there. We start by simplifying the system matrices E_i, F_i, G_i and H_i as described in Lemma 3.3, so that, for $i = 0$ and $i = 1$,

$$E_i \text{ has full column rank} \tag{5.9}$$

$$\begin{pmatrix} zE_i - F_i \end{pmatrix} \text{ has full column rank, } \forall z \tag{5.10}$$

$$\begin{pmatrix} zE_i - F_i & G_i \end{pmatrix} \text{ has full row rank, } \forall z \tag{5.11}$$

$$\begin{pmatrix} E_i & G_i \end{pmatrix} \text{ has full row rank, } \forall z. \tag{5.12}$$

Considering the case of large N. The approximate solution is given by

$$J^i = \sum_{j=1}^{N} \|r_i\|^2 \tag{5.13}$$

where $t_i(0) = 0$ and

$$t_i(j+1) = F_i (E_i^T P_i E_i)^{-1} E_i^T P_i t(j) + H_i \begin{pmatrix} v(j) \\ u(j) \\ y(j) \end{pmatrix} \tag{5.14}$$

$$r_i(j) = W_i t(i) \tag{5.15}$$

and where P_i is the unique positive definite solution of the algebraic descriptor Ricatti equation

$$P_i = \left(F_i \left(E_i^T P_i E_i \right)^{-1} F_i^T + G_i G_i^T \right)^{-1} \tag{5.16}$$

and W_i is any matrix satisfying

$$W_i^T W_i = P_i - P_i E_i (E_i^T P_i E_i)^{-1} E_i^T P_i. \tag{5.17}$$

Systems (5.14)-(5.15), $i = 0, 1$, are two residual generator filters. Their outputs are squared and summed up. And at the end of the test period, the results are compared and a decision is made.

We can also consider a variant of the ratio test:

$$\sum_{k=1}^{m} \|r_i(k)\|^2 > \sigma. \tag{5.18}$$

We decide that a failure has occurred if (5.18) becomes true for $i = 0$, or we decide that no failure has occurred if (5.18) becomes true for $i = 1$, whichever occurs first. This can happen for $m < N$, i.e., a decision can be made before the end of the test period.

6 Conclusion

We have studied the problem of constructing minimum energy input test signals (detection signals) for separating two given models based on input-output measurements. Model and measurement uncertainties are supposed to be bounded energy arbitrary signals. It is shown that as the detection horizon goes to infinity, optimal detection signals converge to pure sinusoids. The ratio of the bound on the norm of the uncertainty signal over the norm of the optimal detection signal plays an important role in our development; we call it the separability index. A constructive method for computing this index, and the corresponding detection signal is given.

A fundamental assumption in this work is that the detection signal is constructed off-line, i.e., the detection signal does not depend on the on-line measurements. It is conceivable to consider a situation where the detection signal depends causally on the measurements. In practice, in most cases, this is not desirable. This dependence introduces a feedback which modifies the dynamic characteristics of the system; it can even destabilize it. It is more realistic to consider the separability index as a control objective to be used in the feedback controller design (assuming that we are dealing with a controlled system). This problem is under investigation.

References

[1] Basseville, M. and Benveniste, A., (eds), Detection of abrupt changes in signals and dynamical systems, *Lecture notes in Control and Information Science*, vol. 77, Springer-Verlag, 1985.

[2] Campbell, S. L. and Meyer, C.D. Jr., *Generalized Inverses of Linear Transformations*, Dover, 1991.

[3] Iglesias, P. A., and Glover, K., State-space approach to discrete-time $\mathcal{H}_\infty$ control, *Int. J. Control*, vol. 54, no. 5, 1991, pp. 1031-1073.

[4] Kailath, T., *Linear Systems*, Prentice Hall, 1980.

[5] Kerestecioğlu, F., Change detection and input design in dynamical systems, *Research Studies Press*, Taunton, U.K., 1993.

[6] Kerestecioğlu, F. and Zarrop, M. B., Input design for detection of abrupt changes in dynamical systems, *Int. J. Control*, vol. 59, no. 4, 1994, pp. 1063-1084.

[7] Kullback, S., *Information Theory and Statistics*, John Wiley & Sons, 1971.

[8] Nikoukhah, R., Guaranteed active failure detection and isolation for linear dynamical systems, *Automatica*, vol. 34, no. 11, 1998.

[9] Nikoukhah, R., Innovations generation in the presence of unknown inputs: application to robust failure detection, *Automatica*, vol. 30, no. 12, 1994.

[10] Patton, R. J., Frank, P. M. and Clark, R. N., (eds), *Fault diagnosis in dynamic systems*, Prentice Hall, 1989.

[11] Uosaki, K, Tanaka, I., and Sugiyama, H., Optimal input design for autoregressive model discrimination with constrained output variance, *IEEE Trans. Auto. Contr.*, vol. AC-29, no. 4, 1984, pp. 348-350.

[12] Van Dooren, P., The computation of Kronecker's Canonical Form of a singular pencil, *Linear Algebra and its Applications*, vol. 27, pp. 103-140, 1979.

[13] Willsky, A. S., A survey of design methods for failure detection in dynamics systems, *Automatica*, vol. 12, pp. 601-611, 1976.

[14] Zhang, X. J., *Auxiliary signal design in fault detection and diagnosis*, Springer-Verlag, Heidelberg, 1989.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399